Structural Estimation by Homotopy Continuation

Philipp Müller
Gregor Reich
University of Zurich

March 25, 2021

Introduction

This paper "started" in the Computational Economics and Finance course taught by Prof. Karl Schmedders in 2017

- Term paper & Semester thesis
  - *Homotopy applied to counterfactual analysis with multiple equilibria. Developing an interface between Fortran & Matlab.*
- Master thesis
  - *Identifiability Analysis In Structural Models Using MPEC And Homotopy Parameter Continuation Methods*

Takeaway from this lecture: insightful research through the combination of various tools you have learned in this course

- Mathematical programming with equilibrium constraints (MPEC)
  - Constrained Optimization
  - Dynamic Programming
- Homotopy Continuation
- Automatic Differentiation

Introduction

Given Two parameterized models (1, 2) with parameters (θ_A, θ_B) and some observed data.

Estimation Find the maximum likelihood estimates of θ_A and θ_B s.t. the model best explains the data.

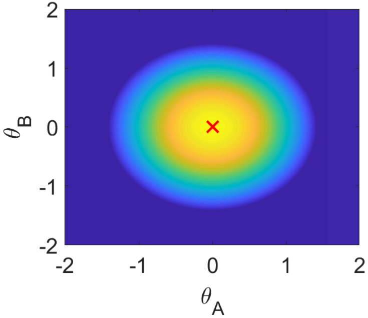


Figure: Model 1: Likelihood.

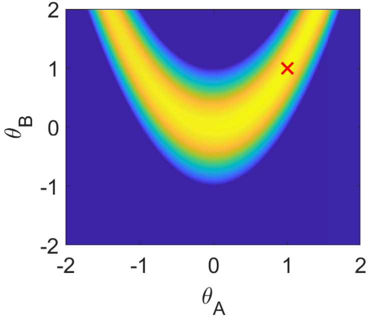


Figure: Model 2: Likelihood.

Introduction

Given Two parameterized models (1, 2) with parameters (θ_A, θ_B) and some observed data.

Estimation Find the maximum likelihood estimates of θ_A and θ_B s.t. the model best explains the data.

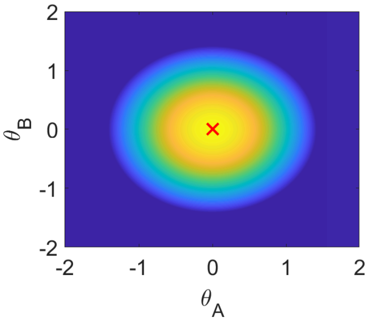


Figure: Model 1: Likelihood.

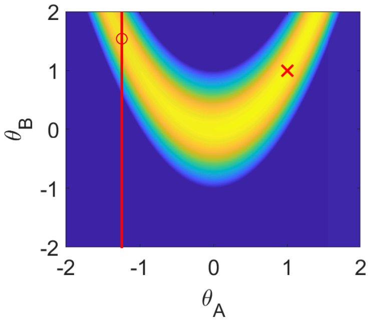


Figure: Model 2: Likelihood.

Introduction

Given Two parameterized models (1, 2) with parameters (θ_A, θ_B) and some observed data.

Estimation Find the maximum likelihood estimates of θ_A and θ_B s.t. the model best explains the data.

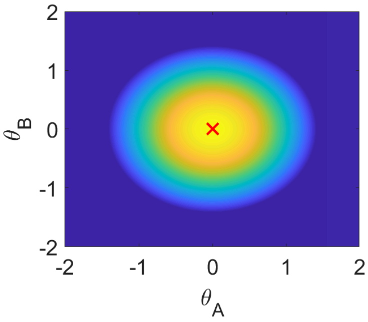


Figure: Model 1: Likelihood.

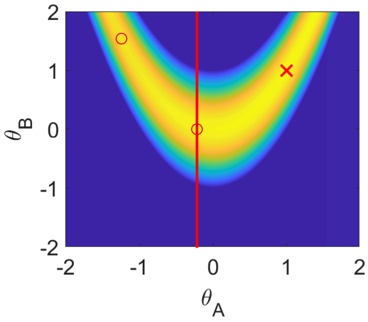


Figure: Model 2: Likelihood.

Introduction

Given Two parameterized models (1, 2) with parameters (θ_A, θ_B) and some observed data.

Estimation Find the maximum likelihood estimates of θ_A and θ_B s.t. the model best explains the data.

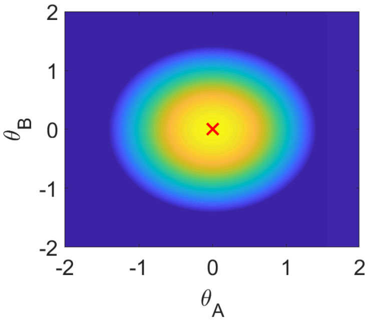


Figure: Model 1: Likelihood.

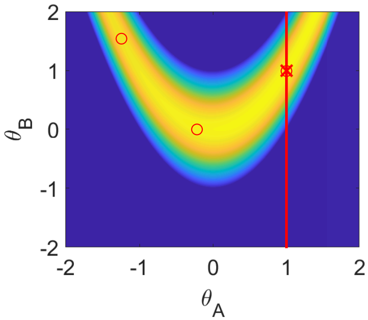


Figure: Model 2: Likelihood.

Introduction

Given Two parameterized models (1, 2) with parameters (θ_A, θ_B) and some observed data.

Estimation Find the maximum likelihood estimates of θ_A and θ_B s.t. the model best explains the data.

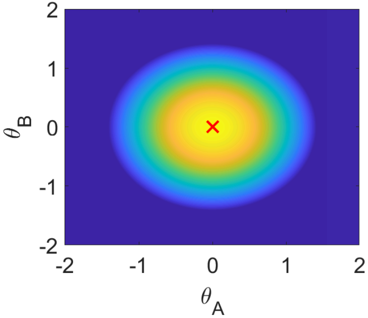


Figure: Model 1: Likelihood.

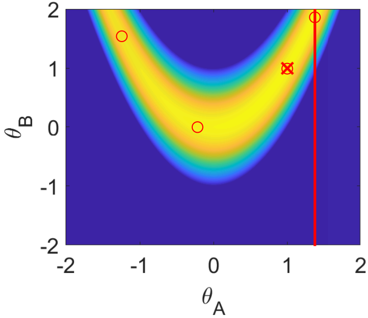


Figure: Model 2: Likelihood.

Introduction

Given Two parameterized models (1, 2) with parameters (θ_A, θ_B) and some observed data.

Estimation Find the maximum likelihood estimates of θ_A and θ_B s.t. the model best explains the data.

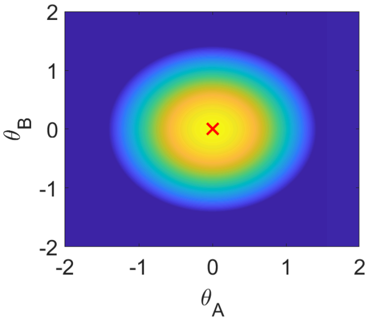


Figure: Model 1: Likelihood.

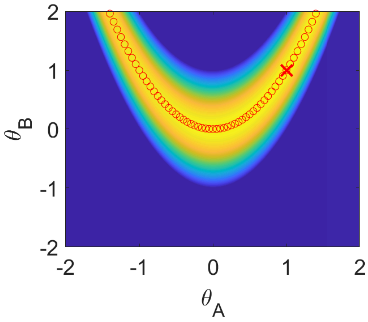


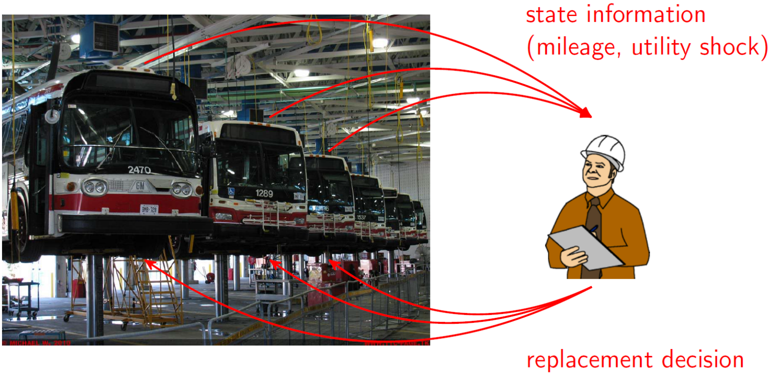
Figure: Model 2: Likelihood.

Introduction

- We focus on dynamic discrete choice models where the discount parameter β is generally considered to be "poorly" identified.
- We propose to formulate
  - the structural estimation as **parameterized** constrained optimization, i.e., parameterized version of Su and Judd [2012]
  - and solve this efficiently by **homotopy parameter continuation**.
- This novel approach enables the econometrician to computationally efficiently
  - estimate the structural parameters **even in models with one poorly identified parameter** and
  - perform inference based on the full (profile) likelihood function (not only point estimates).

The Bus Engine Replacement Model (Rust, 1987)

John Rust: *Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher.* Econometrica, 1987.



Utility Function

- Agent's utility + shock for the single period payoff

$$u(x_t, d_t; \theta_{11}, \mathrm{RC}) + \epsilon_t(d_t) = \begin{cases} -c(x_t, \theta_1) + \epsilon_t(0) & \text{if } d_t = 0 \\ -RC + \epsilon_t(1) & \text{if } d_t = 1 \end{cases}$$

  - $d_t = 0$: performing regular maintenance
  - $d_t = 1$: replacing the engine
- State variables
  - x_t mileage state
  - ϵ i.i.d. gumbel utility shock (only observed by agent)
- **Parameters**
  - θ_{11} regular maintenance cost parameter
  - RC replacement cost parameter

Value Function - Regenerative Optimal Stopping

Objective The agent wants to maximize his expected discounted utility over an **infinite horizon**.

$$V_{\theta,\beta}(x_t, \epsilon_t) = \max_{D(x_t)\in\mathcal{D}} \mathbb{E}\left[\sum_{j=t}^{\infty} \beta^{j-t} \left(u(x_j, D(x_j); \theta_1, \text{RC}) + \epsilon(D(x_j))\right) \Big| x_t\right]$$

where $\theta \equiv (\text{RC}, \theta_1)$ and $D(\cdot)$ denotes the policy function.

Bellman $V_{\theta,\beta}$ is the unique solution to the Bellman equation

$$V_{\theta,\beta}(x, \epsilon) = \max_{d\in\{0,1\}} [u(x, d, \theta_1) + \epsilon(d) + \beta\mathbb{E}[V_{\theta,\beta}(x', \epsilon')|x, d]],$$

where x' and ϵ' denote the next period state variables.

Preference $\beta = 1$: Maximize the long-run average utility [Bertsekas, 2012].

$\beta > 1$: Maximize today's and future utility - "future-bias" [Blom Västberg and Karlström, 2017].

Challenges Arising for $\beta \to 1$ and $\beta > 1$

- Let's have a look at the definition of the value function

$$V_{\theta,\beta}(x_t, \epsilon_t) = \max_{D(x_t)\in\mathcal{D}} \mathbb{E}\left[\sum_{j=t}^{\infty} \beta^{j-t} \left(u(x_j, D(x_j); \theta_1, \mathrm{RC}) + \epsilon(D(x_j))\right) \Big| x_t\right]$$

⚠ $\overline{V} \to -\infty$ for $\beta \to 1$

- The classic value iteration solves for the discretized (expected) value by

$$\overline{V} = T_{\theta,\beta}(\overline{V}),$$

with $T_{\theta,\beta}(\cdot)$ denoting the Bellman operator.

⚠ The contraction mapping fails for $\beta \geq 1$. Note that MPEC does not rely on the contraction mapping property.

Challenges Arising for $\beta \to 1$ and $\beta > 1$

- Let's have a look at the definition of the value function

$$V_{\theta,\beta}(x_t, \epsilon_t) = \max_{D(x_t)\in\mathcal{D}} \mathbb{E}\left[\sum_{j=t}^{\infty} \beta^{j-t} \left(u(x_j, D(x_j); \theta_1, \text{RC}) + \epsilon(D(x_j))\right) \middle| x_t\right]$$

⚠ $\overline{V} \to -\infty$ for $\beta \to 1$

- The classic value iteration solves for the discretized (expected) value by

$$\overline{V} = T_{\theta,\beta}(\overline{V}),$$

with $T_{\theta,\beta}(\cdot)$ denoting the Bellman operator.

⚠ The contraction mapping fails for $\beta \geq 1$. Note that MPEC does not rely on the contraction mapping property.

Relative Value Iteration

Recall We can solve for the optimal value function by solving the fixed-point equation $\overline{V} = T_{\theta,\beta}(\overline{V})$.

Note Even though, the value function $V \to \infty$, the **difference** between the value at different states might be finite

- Equation (1) solves for the **relative** value function $h = \overline{V} - \overline{V}_1$ as

$$h = T_{\theta,\beta}(h) - T_{\theta,\beta}(h)_1 \tag{1}$$

with $h \in \mathbb{R}^{90}$ Bertsekas [2012] .

Structural Estimation

Objective Identify the most likely values for the parameters $\theta = (\theta_{11}, \mathrm{RC})$ and β given the observed data.

Data $\sim$ 8000 observations of the state and control variables (= mileage states and replacement decisions).

Approach Simultaneously solve the **likelihood** and **fixed-point** problem

$$\theta^*, \beta^* = \arg\max_{\theta,\beta} L(h, \theta, \beta; \{x_t, d_t\})$$

$$h = T_{\theta,\beta}(h) - T_{\theta,\beta}(h)_1|_{\theta=\theta^*,\beta=\beta^*}$$

- Two popular solution methods are the nested fixed-point algorithm (NFXP) Rust [1987] and the mathematical programming with equilibrium constraints by Su and Judd [2012]

MPEC

- Su and Judd [2012] formulate the structural estimation as constrained optimization

$$\max_{(h,\theta,\beta)} L(\theta, \beta, h; \{x_t, d_t\}),$$
$$\text{s.t.} \; h = T_{\theta,\beta}(h) - T_{\theta,\beta}(h)_1,$$

with $\theta \in \mathbb{R}^2$, $\beta \in \mathbb{R}_+$, and $h \in \mathbb{R}^{90}$.

MPEC

- Su and Judd [2012] formulate the structural estimation as constrained optimization

$$\max_{(h,\theta,\beta)} L(\theta, \beta, h; \{x_t, d_t\}),$$
$$\text{s.t. } h = T_{\theta,\beta}(h) - T_{\theta,\beta}(h)_1,$$

with $\theta \in \mathbb{R}^2$, $\beta \in \mathbb{R}_+$, and $h \in \mathbb{R}^{90}$.

- If the likelihood is (almost) flat w.r.t. changes in β, we refer to β as "poorly identified". Numerically, its Hessian becomes (nearly) singular and in turn, the maximum likelihood estimation becomes numerically hard.

MPEC

- Su and Judd [2012] formulate the structural estimation as constrained optimization

$$\max_{(h,\theta,\cancel{\beta})} L(\theta, \beta, h; \{x_t, d_t\}),$$
$$\text{s.t. } h = T_{\theta,\beta}(h) - T_{\theta,\beta}(h)_1,$$

with $\theta \in \mathbb{R}^2$, $\beta \in \mathbb{R}_+$, and $h \in \mathbb{R}^{90}$.

- We refer to β as "poorly identified" if the likelihood is (almost) flat w.r.t. changes in β. Numerically, its Hessian becomes (nearly) singular and in turn, the maximum likelihood estimation becomes numerically hard.
- Thus, β is often **calibrated** to some value.

Profile Likelihood

- Instead of calibrating β to some value, we propose to solve for the maximum likelihood estimates **as a function** of the *controlled parameter*,
  i.e., as **parametric** maximum likelihood estimates
- By setting β as controlled parameter we define

$$L_p(\beta) = \max_{\theta,h} L(\theta, h; \{x_t, d_t\}, \beta)$$
$$\text{s.t. } h = T_{\theta,\beta}(h) - T_{\theta,\beta}(h)_1,$$

  i.e., we optimize w.r.t. all parameters **but** the controlled parameter β
- This is equivalent to the notion of a *profile likelihood*.

First-Order Necessary Optimality Conditions

- Its Lagrangian $\mathcal{L}$ is defined as

$$\mathcal{L}(\theta, h, \boldsymbol{\mu}) = L(\theta, h; \beta) - \sum_i \mu_i \left(h - T_{\theta,\beta}(h) + T_{\theta,\beta}(h)_1\right)$$

- If (θ^*, h^*) is a local optimal solution to $L_p(\beta)$, where the LICQ holds, then there exists a unique μ^* s.t.

$$\nabla_{(\theta,h,\mu)}\mathcal{L}(\theta^*, h^*, \mu^*; \beta) = 0. \tag{2}$$

PMPEC - Summary

- The structural estimation as parametric constrained optimization

$$\max_{(h,\theta)} L(\theta, \beta, h; \{x_t, d_t\}),$$
$$\text{s.t. } h = T_{\theta,\beta}(h) - T_{\theta,\beta}(h)_1,$$

- First-order necessary conditions (Lagrange) form a parametric system of equations

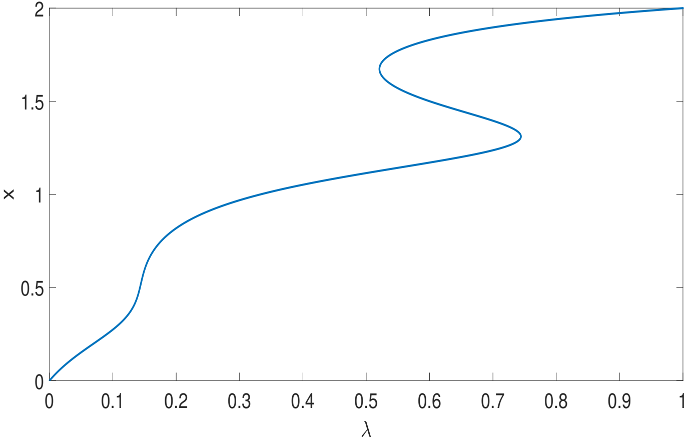
$$\nabla_{(\theta,h,\mu)} \mathcal{L}(\theta^*, h^*, \mu^*; \beta) = 0.$$

- We are interested in the solution manifold of the parametrized FOC (profile likelihood as implicit function)

$$c \equiv \{(\beta, \theta, h, \mu) : \nabla_{\theta,h,\mu} \mathcal{L}(\beta, \theta, h, \mu) = 0\}$$

Recap Homotopy Continuation

$$H(x, \lambda) = x + \lambda(x - 4 + \sin(2\pi x))$$



Recap Homotopy Continuation: Algorithm

Objective Find the solution to $H(x, \lambda) = 0$ *for all* $\lambda \in [0, 1]$ by tracing the curve $c := \{(x, \lambda) : H(x(s), \lambda(s)) = 0\}$ with arclength s.

- We have a starting point. To stay on the zero curve as we "move along":

$$\frac{\partial H(x(s), \lambda(s))}{\partial s} = 0 \tag{3}$$

- The initial and boundary value problem (IBVP) reads

$$\frac{\partial H(x(s), \lambda(s))}{\partial x} x'(s) + \frac{\partial H(x(s), \lambda(s))}{\partial \lambda} \lambda'(s) = 0 \tag{4}$$

$$x(0) = x_0, \quad \lambda(0) = 0, \quad ||(x'(s), \lambda'(s))||_2^2 = 1, \tag{5}$$

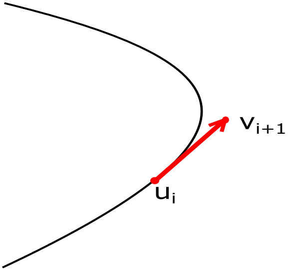
Recap Homotopy Continuation: Algorithm

Approach Trace c by *alternating* **prediction** and **correction** steps.

Predictor Use e.g., Euler's explicit step to predict

$$v_{i+1} = u_i + h \cdot H'(x(s_i), \lambda(s_i)).$$

Corrector Use the predicted point v_{i+1} and improve prediction by e.g., Newton-type methods.



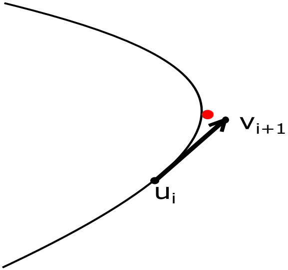
Recap Homotopy Continuation: Algorithm

Approach Trace c by *alternating* **prediction** and **correction** steps.

Predictor Use e.g., Euler's explicit step to predict

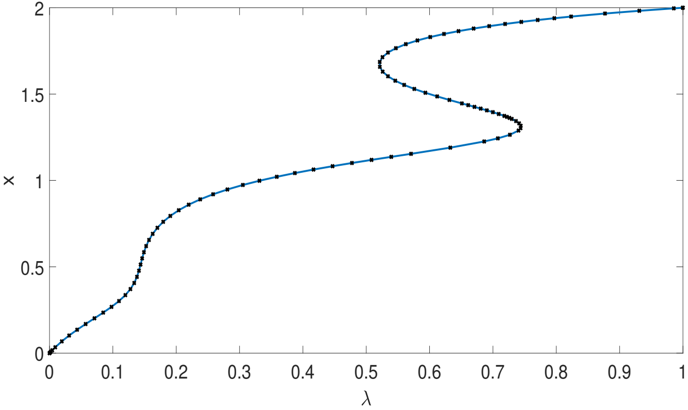
$$v_{i+1} = u_i + h \cdot H'(x(s_i), \lambda(s_i)).$$

Corrector Use the predicted point v_{i+1} and improve prediction by e.g., Newton-type methods.



Recap Homotopy Continuation: ODE-based Algorithm

$$H(x, \lambda) = x + \lambda(x - 4 + \sin(2\pi x))$$



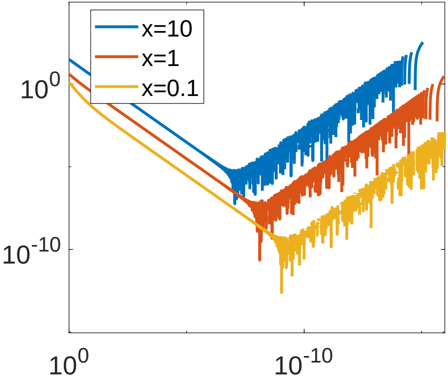
Source: M-Hompack

Recap Finite Differences: Accuracy

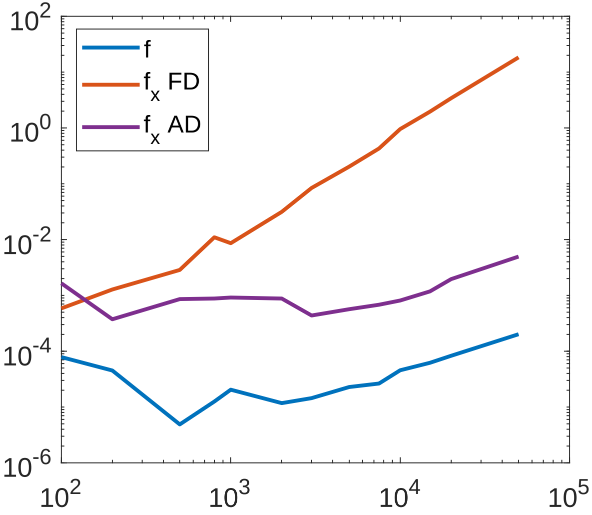
Apply forward differences to

$$f(x) = x^3$$

and decrease step size from 1 to 10^{-16}



Recap Automatic Differentiation: Scaling

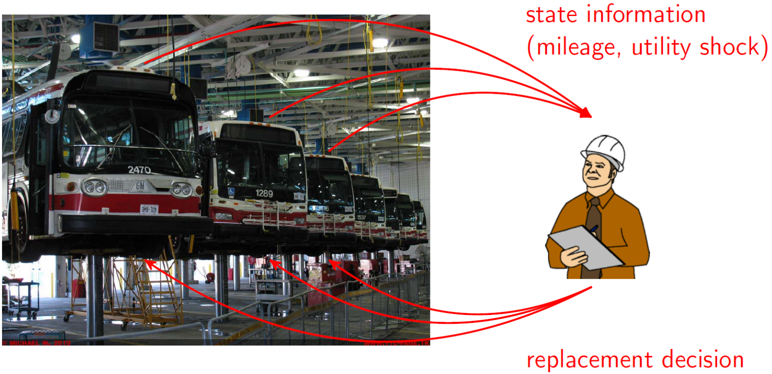


Software Tools

- Homotopy parameter continuation
  - HOMPACK90 by Watson et al. [1997]: a Fortran 90 collection of homotopy solution methods
  - M-Hompack by Müller and Reich [2018]: an interface between Matlab and HOMPACK90 to easily access and employ the efficient homotopy solution methods
- Automatic Differentiation
  - Especially for the homotopy continuation, fast and accurate derivatives are mandatory
  - AD provides analytic derivatives by source code transformation via successively utilizing the chain rule. We use CasADi by Andersson et al. [2018].

The Bus Engine Replacement Model (Rust, 1987)

John Rust: *Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher.* Econometrica, 1987.

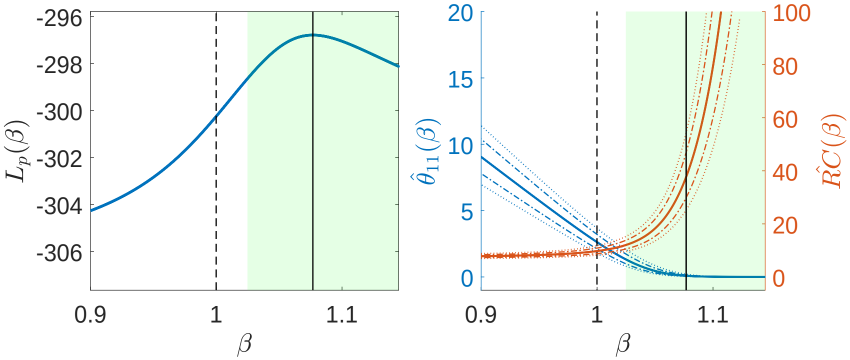


Rust [1987] on the Discount Factor β

"not able to precisely estimate the discount factor β [...]

Changing β to .98 or .9999 produced negligible changes in the likelihood function and parameter estimates [...]

I did note a systematic tendency for the estimated value of β to be driven to 1." Rust [1987]



	β	RC	θ_{11}	L
Rust (1987)	0.9999	9.7558	2.6275	-6055.250
	-	[8.200, 11.76]	[1.810, 3.669]	
MR	1.0768	37.7109	0.0905	-6051.792
	$[1.025, \infty)$	[13.00, 354.9]	[0.001, 1.029]	

Rust [1987]'s Assumed Value for β and Likelihood

"not able to precisely estimate the discount factor β [...]

Changing β to .98 or .9999 produced negligible changes in the likelihood function and parameter estimates [...]

I did note a systematic tendency for the estimated value of β to be **driven to 1**." Rust [1987]

$\Rightarrow$ We numerically show that β is identified with $\hat{\beta} = 1.0768$. Note that the relative value iteration is essential to derive this result! Rust used the value function which diverges for $\beta \to 1$.

The Macroeconomic Historical Context

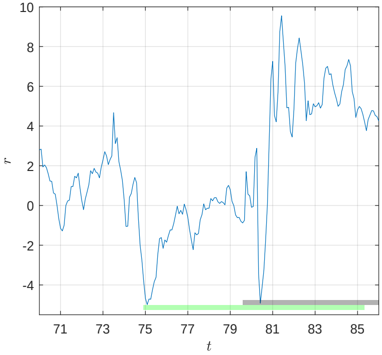


Figure: Blue line: Real interest rate in the US. Grey: Paul Volcker took office as chairman of the Federal Reserve. Green: Rust [1987] dataset.

Structural Break in the Discount Factor

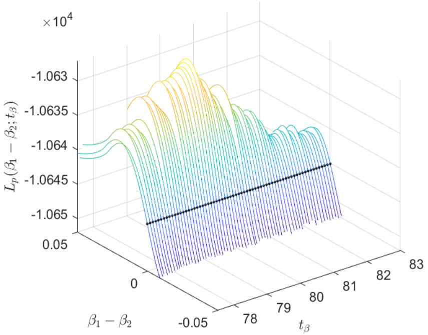


Figure: Homotopy path tracing for each possible time of the structural break, t_β. Black dots denote the restricted model without structural break in the discount factor.

Structural Break in the Discount Factor

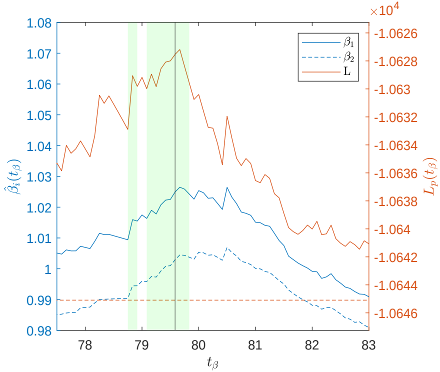


Figure: Profile likelihood $L_p(t_\beta)$; Discount factor estimate before the structural break , $\hat{\beta}_1$, and after the structural break, $\hat{\beta}_2$. Black vertical line denotes the month Paul Volcker took office. Orange dashed line the likelihood of the restricted model, i.e., $\beta_1 = \beta_2$.

Discussion: Identification

" In most applications [the discount factor] is not estimated because it is poorly identified (e.g., see Rust, 1987)." **?**

$\Rightarrow$ Maybe? Many DDCM models we found base on proprietary data; hence, a replication is not possible.

Discussion: $\beta > 1$

"The intertemporal factor is usually assumed to be between 0 and 1 because it is assumed to be 1 / (1 + interest rate) although behaviorally this does not have to hold." **?**

$\Rightarrow$ Note that the decision maker is Harold Zurcher, who, behaviorally, might well have a $\beta > 1$

"We first estimated the discount factor as a parameter, and it turned out to equal 1.001." **?**

Think of yourself and your behavior when paying bills: Would you rather pay them immediately or postpone it?

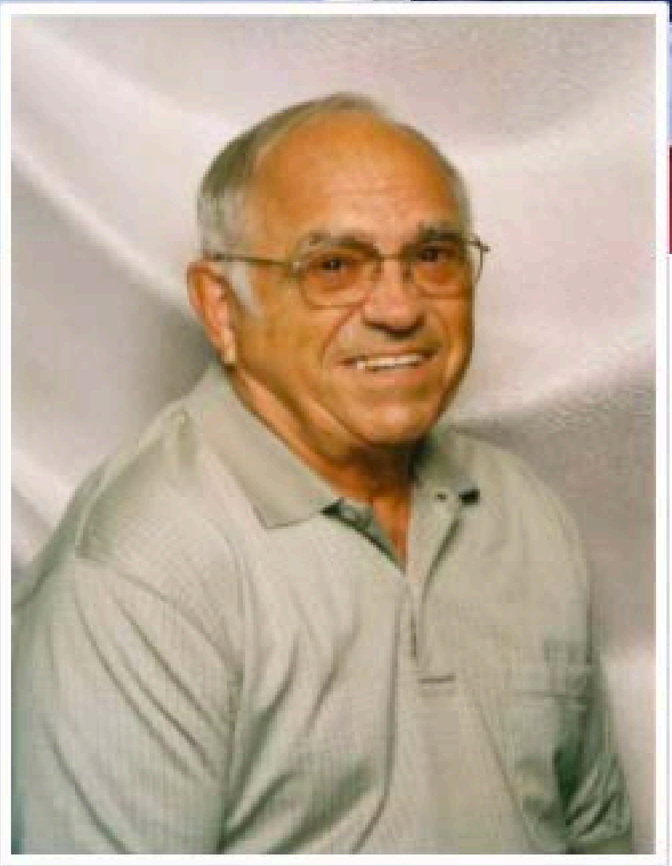
Individuals are not infinitely lived

The infinite-horizon decision making model does not hold for individuals
The infinite-horizon model is a stationary simplification of the true finite-horizon model

Turnpike idea: Decisions today are insensitive to value function at a terminal time in the distant future.

Finite-horizon value functions exist for any β.

The infinite-horizon value function is unimportant
The relativized infinite-horizon value function contains information about decision rules



Harold Alois Zuercher

June 16, 1926 ~ June 21, 2020 (age 94)

Obituary & Services

Robustness: Heteroscedasticity

- Is it all just misspecification?
- The utility shock ϵ is *i.i.d.* EV1 distributed for all mileage states
  - $\Rightarrow$ buses with high mileages are modeled with the same shock distribution as buses with low mileages
  - Fair to assume that the risk of a bad shock increases with mileage
- We check for heteroscedasticity by adding a second shock $\eta \sim N(0,1)$ to the utility function $u + \epsilon$. The resulting utility reads

$$u(x; \theta, d) + \epsilon(d) + \theta_h x \eta \tag{6}$$

where θ_h denotes the heteroscedasticity parameter.

Robustness: Heteroscedasticity

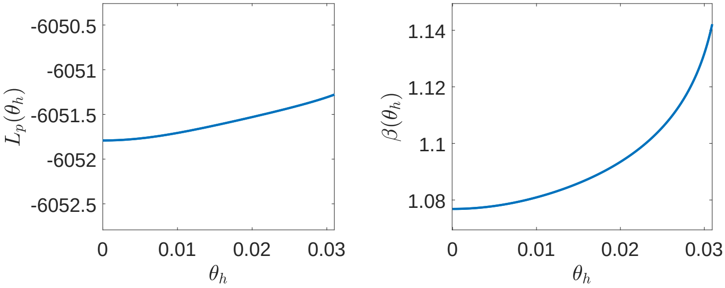
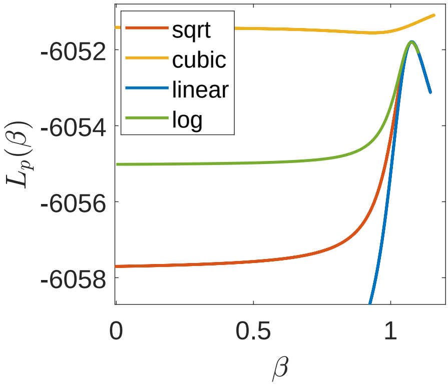


Figure: Heteroscedasticity plots for tracing the heteroscedasticity θ_h from the standard model ($\theta_h = 0$) to ($\theta_h = 0.031$). The left plot depicts the corresponding maximum likelihood estimate $\beta(\theta_h)$ and the right plot depicts $\mathcal{L}_p(\theta_h)$.

Opposed to the intuition, $\beta(\theta_h)$ is monotonically increasing

Robustness: Cost functions



$$u(x_t, d_t; \theta_{11}, \mathrm{RC}) + \epsilon_t(d_t) = \begin{cases} -c(x_t, \theta_1) + \epsilon_t(0) & \text{if } d_t = 0 \\ -RC + \epsilon_t(1) & \text{if } d_t = 1 \end{cases}$$

Conclusion

Method

- Capable of systematic and efficient structural estimation, even for models with a poorly identified parameter and in the presence of multiple equilibria.
- We enable inference on the full (profile) likelihood function.

Model

- Given the original data set and model we can reject that β is unidentified.
- The estimate for β is unexpectedly even statistically significantly larger than 1 with $\beta = 1.078$ ($p = 0.0086$).

References I

J. A. E. Andersson, J. Gillis, G. Horn, J. B. Rawlings, and M. Diehl. CasADi – A software framework for nonlinear optimization and optimal control. *forthcoming in: Mathematical Programming Computation*, 2018.

D. P. Bertsekas. *Dynamic Programming and Optimal Control* , volume 2 of *Approximate dynamic programming*. Athena Scientific, Belmont, MA, 4th edition, 2012.

O. Blom Västberg and A. Karlström. Discount factors greater than or equal to one in infinite horizon dynamic discrete choice models. *unpublished; available on request from the authors*, 2017.

J. Rust. Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher. *Econometrica: Journal of the Econometric Society*, 55(5):999–1033, 1987.

References II

C.-L. Su and K. L. Judd. Constrained Optimization Approaches to Estimation of Structural Models. *Econometrica: Journal of the Econometric Society*, 80(5):2213–2230, 2012.

L. T. Watson, M. Sosonkina, R. C. Melville, A. P. Morgan, and H. F. Walker. Algorithm 777: HOMPACK90: a suite of Fortran 90 codes for globally convergent homotopy algorithms. *ACM Transactions on Mathematical Software (TOMS)*, 23(4): 514–549, Dec. 1997.

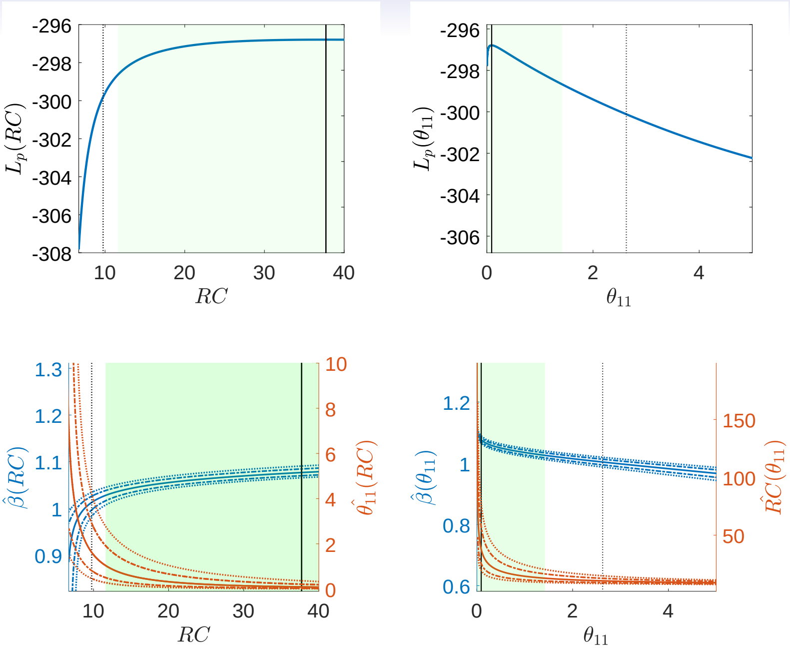
Confidence Intervals

The γ-*likelihood ratio confidence interval* of parameter θ_j as function of β reads

$$\left\{\theta_j : \max_{\theta_{-j}} L(\theta;\beta) - \Big(L(\hat{\theta}(\beta);\beta) - 0.5\chi_1^2(\gamma)\Big) \geq 0\right\}, \qquad (7)$$

$\hat{\theta}(\beta)$ denotes the maximum likelihood estimate in dependence of β. This naturally integrates into our tracing approach

$$\begin{pmatrix} L(\theta;\beta) - (L(\hat{\theta}(\beta);\beta) - 0.5\chi_1^2(\gamma)) \\ \nabla_{\mu,\theta_{-j},\sigma}\mathcal{L}(\mu,\theta,\sigma;\beta) \end{pmatrix} = 0. \qquad (8)$$



-296
-298
-300
-302
-304
-306
-308
$L_p(RC)$
10
20
30
40
RC
$L_p(\theta_{11})$
0
2
4
θ_{11}
1.3
1.2
1.1
1
0.9
$\hat{\beta}(RC)$
10
8
6
4
2
0
$\hat{\theta_{11}}(RC)$
0.8
0.6
$\hat{\beta}(\theta_{11})$
150
100
50
$\hat{RC}(\theta_{11})$